Yawn, Linux for HPC, What Next?

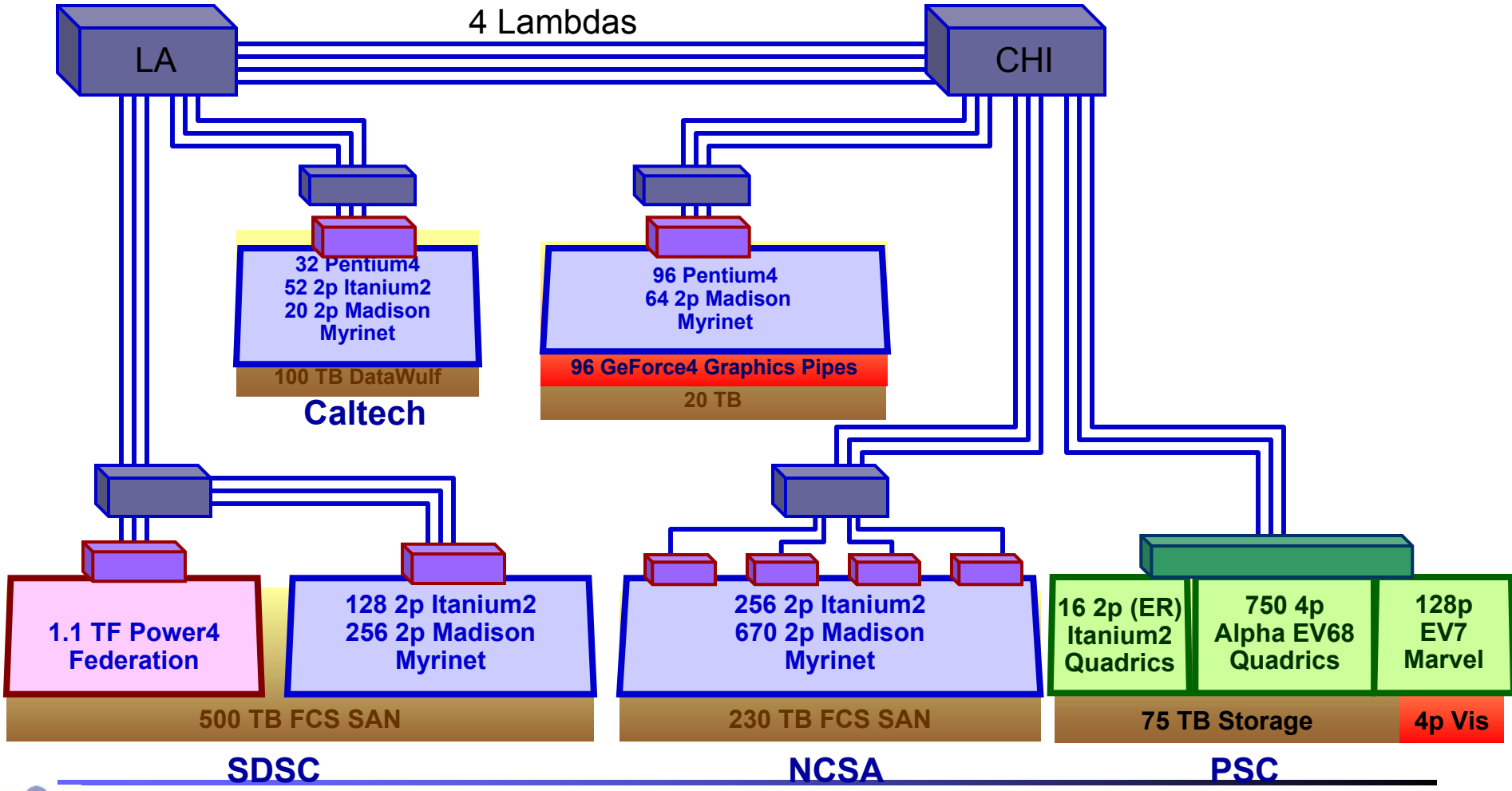
Pete Beckman

Argonne National Laboratory

The Evolution of Linux for HPC

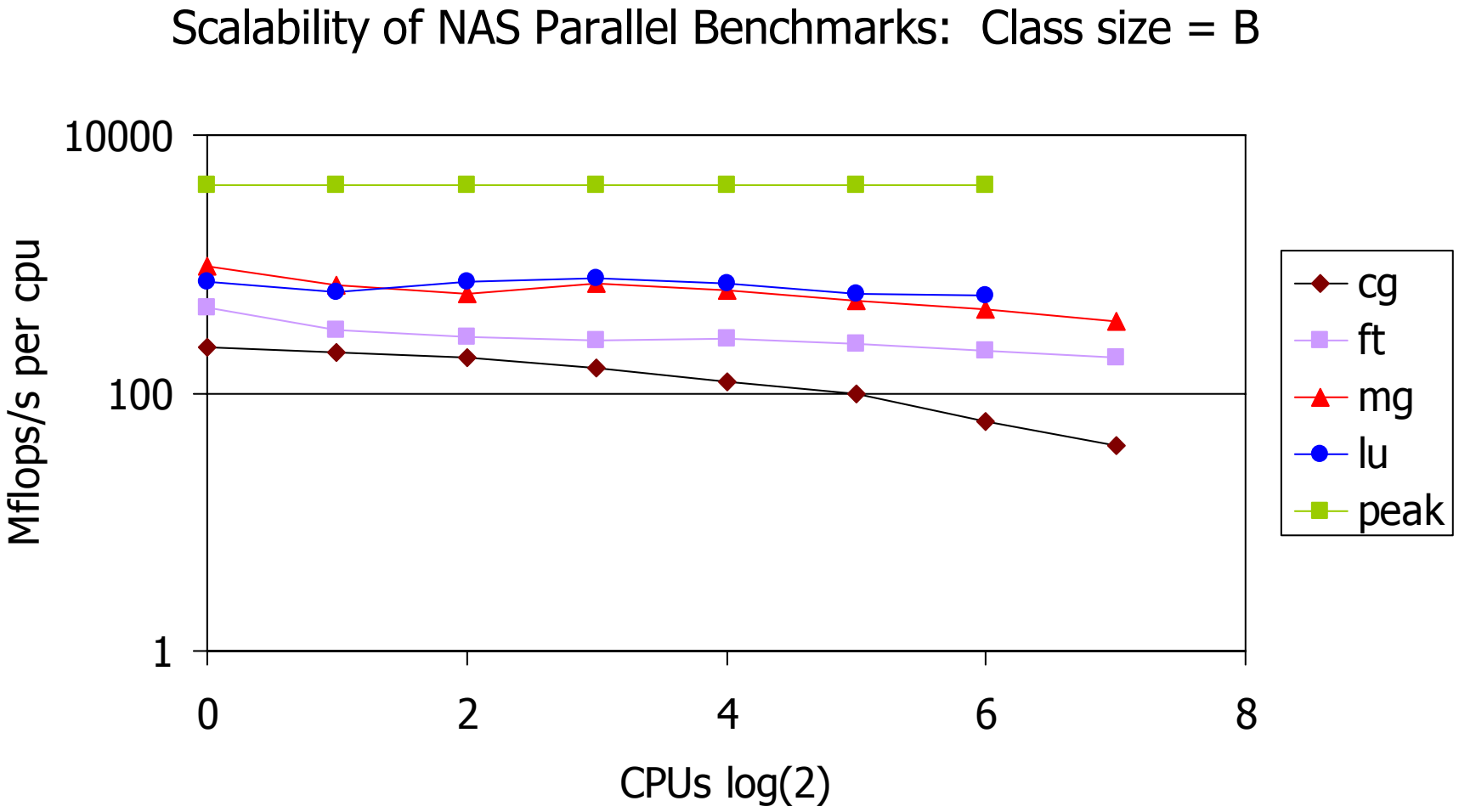
- **1993**: Can a Linux cluster provide cost effective cycles for a single scientist?
- **1997**: Linux cluster wins Gordon Bell award for cost/performance
- **1998**: Linux cluster makes Top500, the press calls it a "Supercomputer"
- **2000**: Can a Linux cluster provide a stable, multi-user HPC environment for a wide-variety of applications?
- **2003**: Linux clusters provide DOE & NSF centers production environments for thousands of users

TeraGrid Deployment, Fall 2003



Source: Rob Pennington <robp@ncsa.uiuc.edu>

NAS Parallel Benchmarks on TeraGrid McKinley

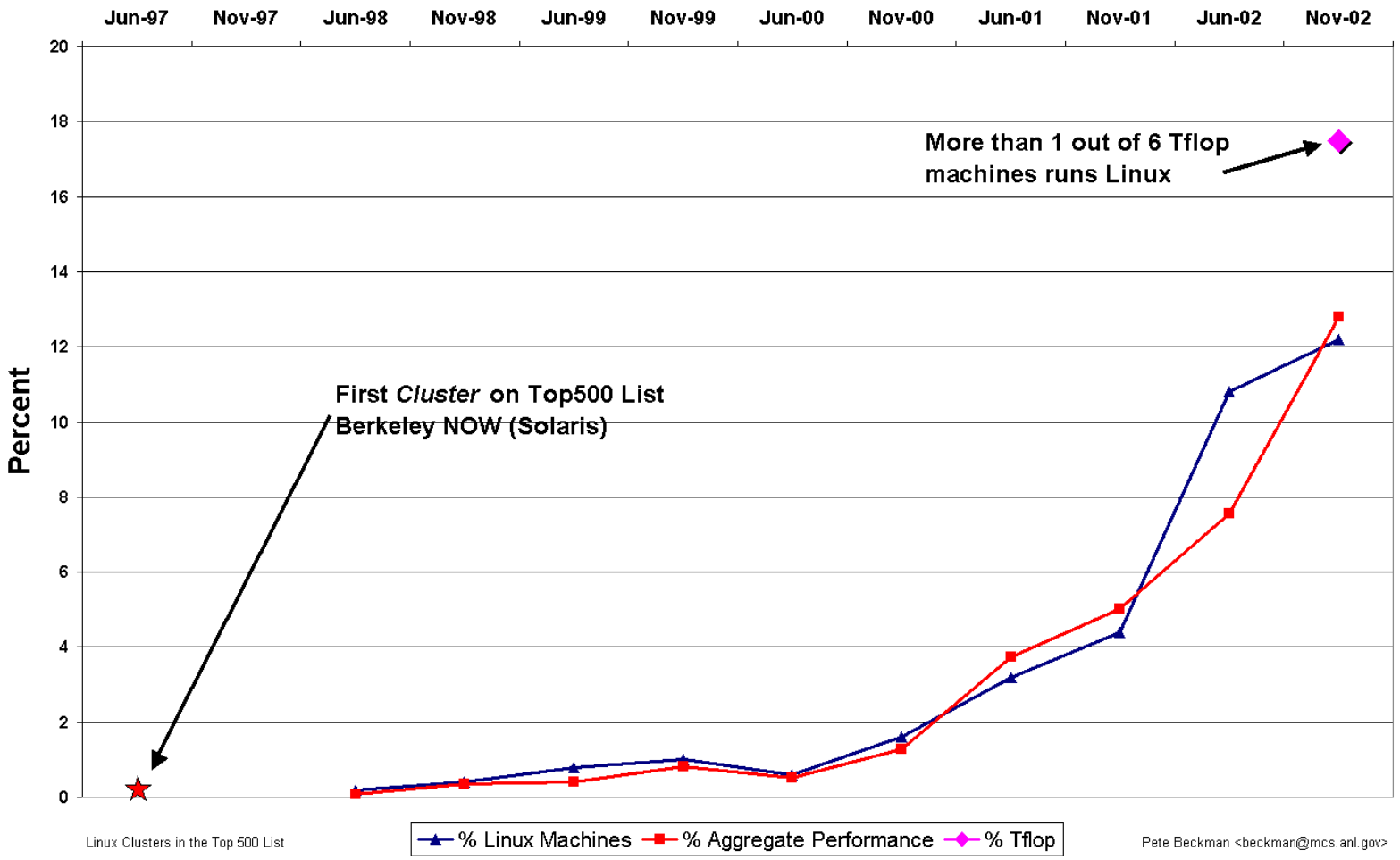


Source: Sharon Brunett <sharon@cacr.caltech.edu>

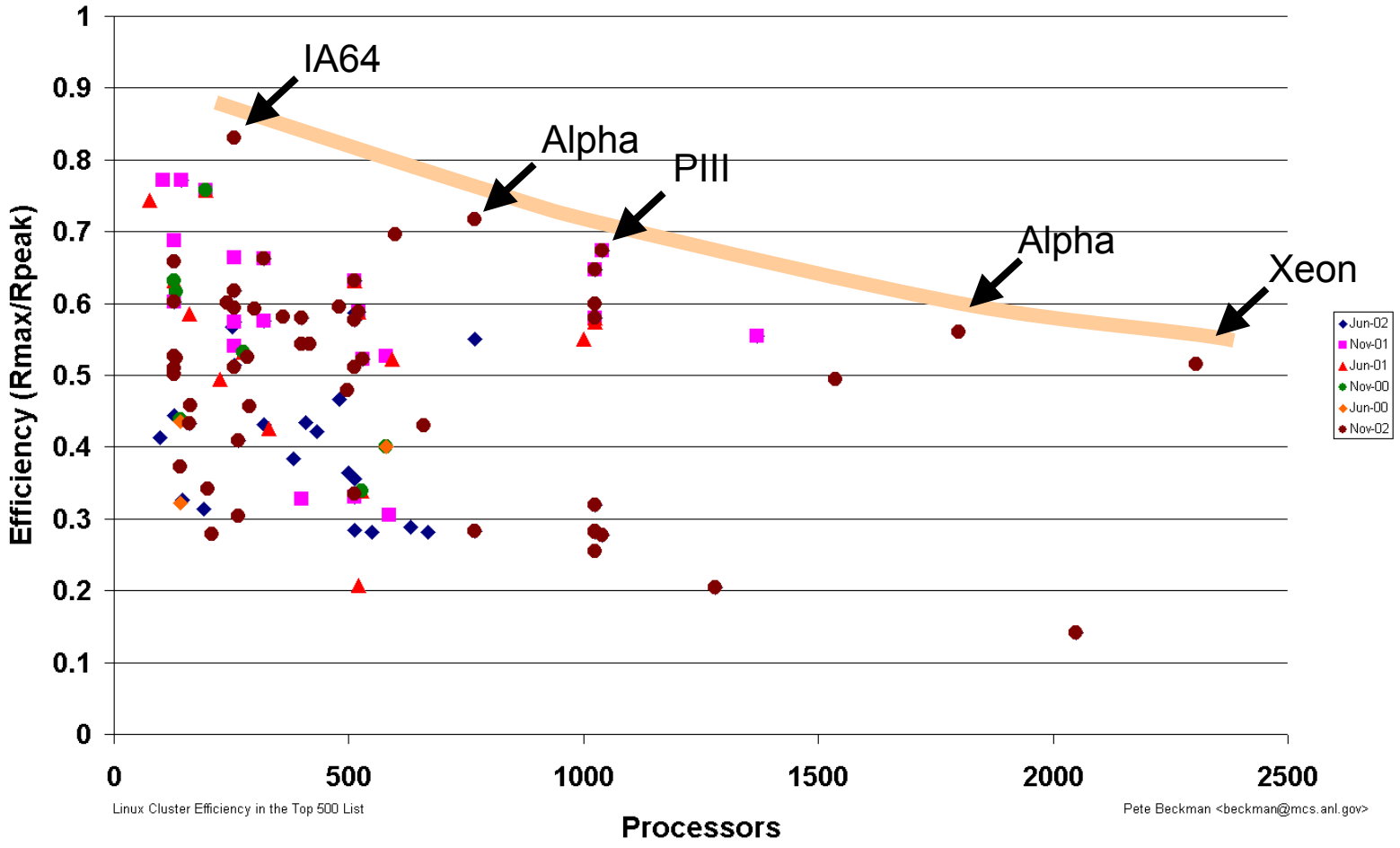
Exploring The Next Question:

- **2003**: Will Linux become the **ONLY** environment for HPC?

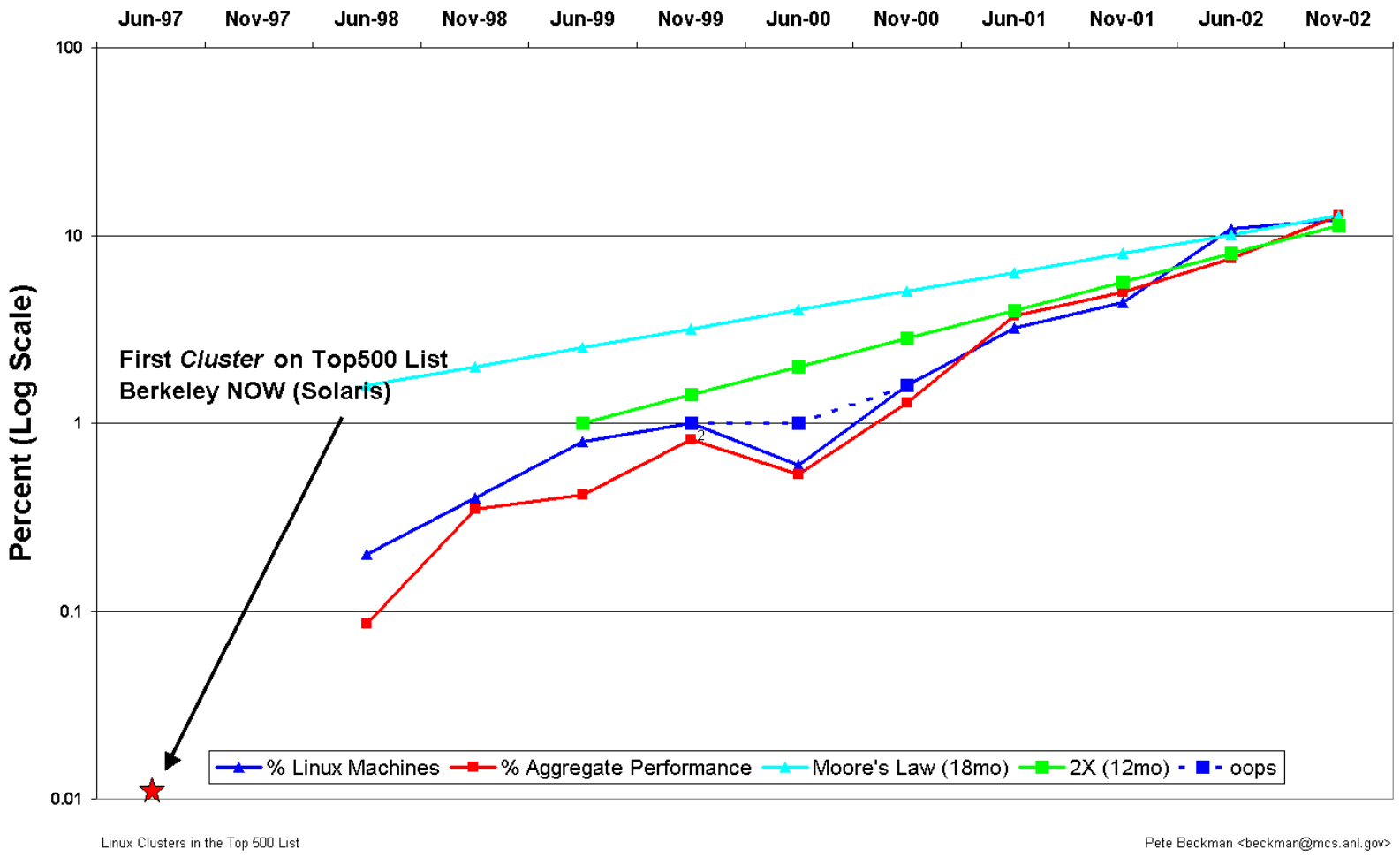
Linux Clusters in the Top 500



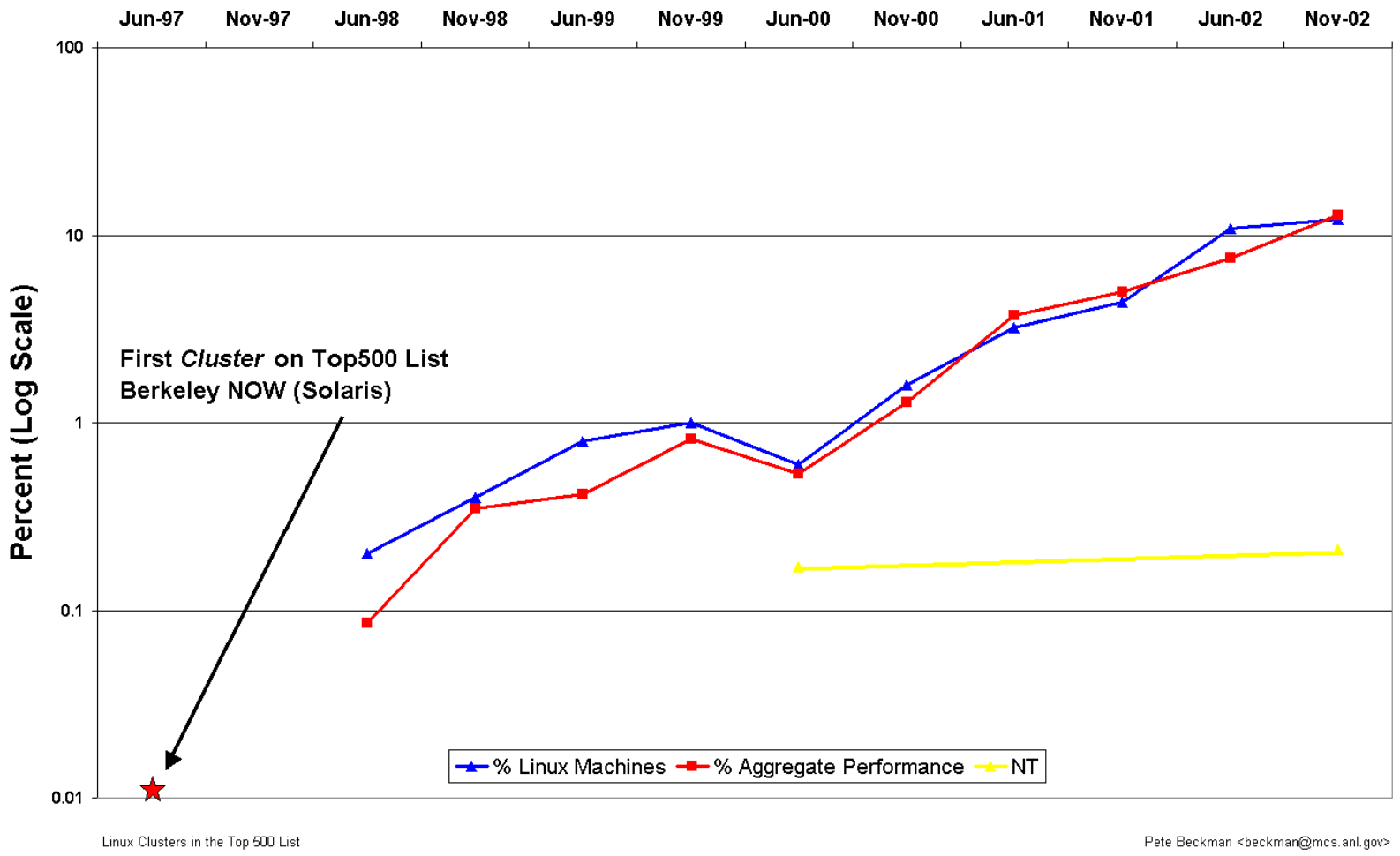
Linux Efficiency in the Top 500



Linux: Plotting The Future

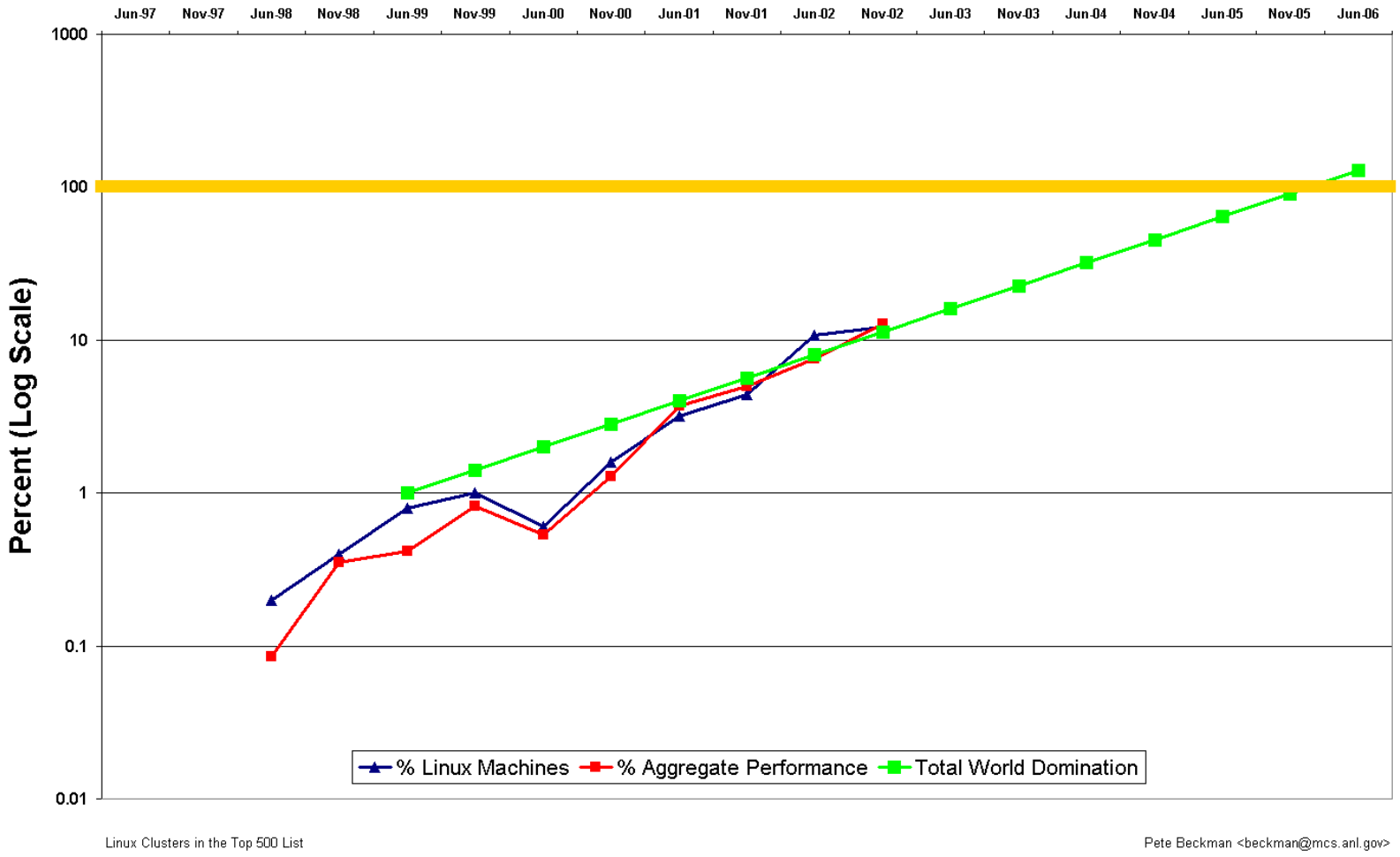


Building Top 500 Clusters with NT



Predicting Future Market Share

How Long Until Total World Domination?



Observations

- The adoption rate of Linux HPC is phenomenal!
  - Linux in the Top500 is doubling every 12 months
  - Linux adoption is not driven by bottom feeders
    - ***Adoption is actually faster at the ultra-scale!***
- Prediction: by 2005, most top-performing supercomputers will be running Linux
- Adoption rate driven largely by economics and human factors
- Possible barriers to continued adoption
  - Better cost/perf of another solution
  - "capability" of total perf or key component is poor

Dominating Market Share

- Linux Clusters: Extremely low barrier to entry
- Effort to learn programming paradigm, libs, devl env., and tools preserved across many orders of magnitude
- It is what engineering and CS students are learning NOW!
  - Example:

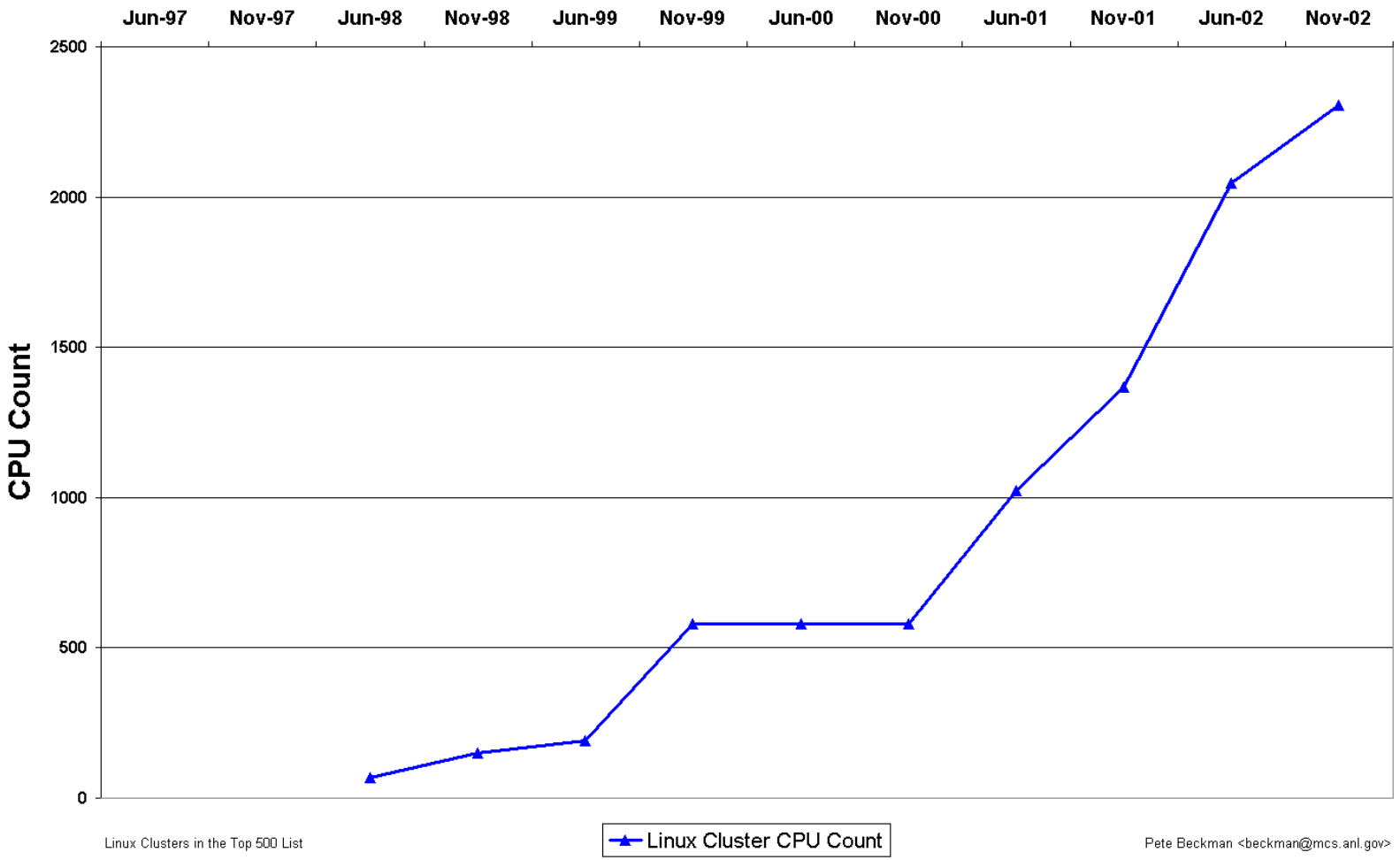
CS 290: Cluster Computing and the Grid

In this class you will learn how to architect an effective cluster, how to properly configure it with software tools (parallel messaging libraries, batch queuing systems, parallel file systems), and understand its performance limitations. […]

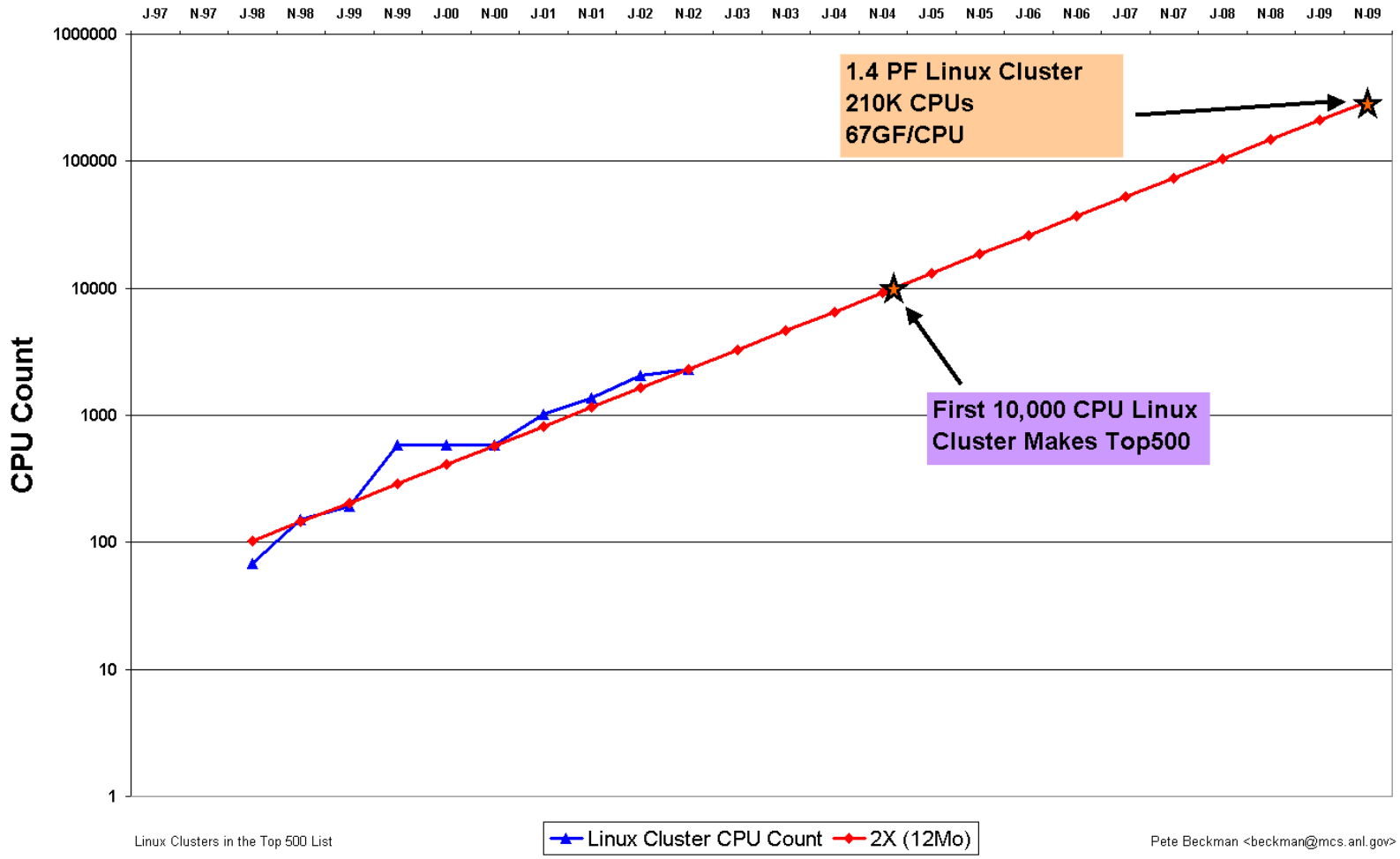
Exploring The Next Question:

- **2003**: How large can Linux clusters get?

Linux Cluster Sizes in the Top 500



Linux Cluster Sizes: Plotting The Future



Observations

- The CPU counts for the largest Linux clusters are currently doubling every year
- Prediction: by 2005, we will have a 10,000 CPU Linux cluster
- Possible barriers
  - (not a market share question)
  - Hmmmmmm…

Possible Barriers and Walls (1 of 3)

- Interconnects
  - Cost models for 10K – 100K node interconnects would suggest full bisection scaling difficult
  - Density of packaging, cost, and scale may encourage multiple, hierarchical interconnects
    - CPU to CPU links
    - Fast copper in cabinet
    - Fiber between cabinets
- RAS
  - Current Linux clusters take very little advantage of RAS features in hardware, little if any real Open Source software exists in this domain

Possible Barriers and Walls (2 of 3)

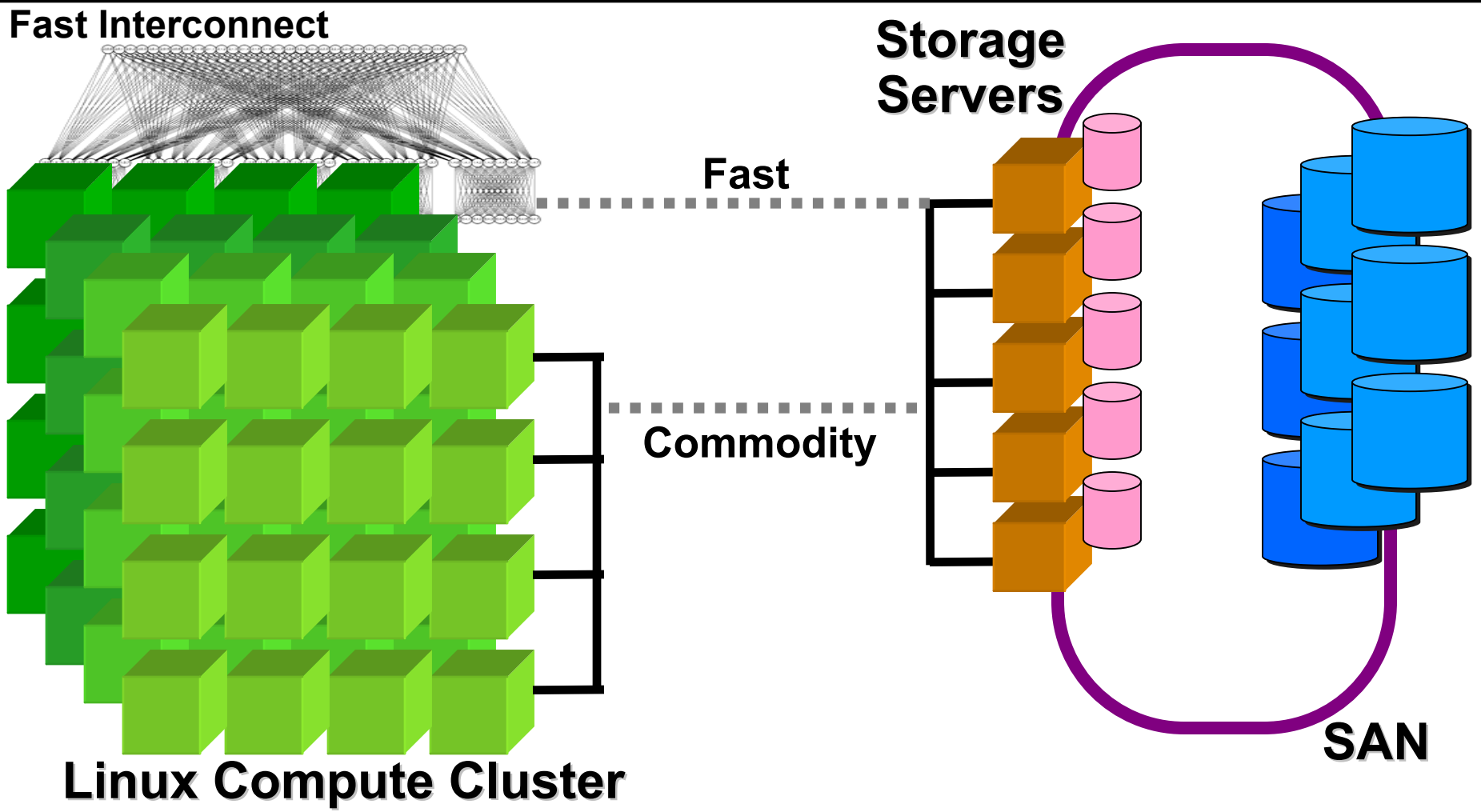
- Fault Tolerance
  - Google claims 54K nodes, what do they do when one goes down?
  - We must do better than "restart ALL nodes from last chkpt"
- System software & mgmt

  for (i=0, i<100000,i++) { … }

Classic Mythology 101 (for Computing)

- Automatic Parallelizing Compiler
- Scalable Distributed Shared Memory
- High Performance Java
- Real-Time Computational Steering
- Write once, run well on any architecture
- A Parallel, Distributed, Global, Cluster, High-Performance, Fault Tolerant, Home Directory File System

Parallel File Systems



Data Access Patterns For File I/O

- Throughput I/O (homedirs)
  - thousands of clients, large file server(s), highest possible aggregate throughput & availability
    - On independent files, the server supports a max of <x> GB/sec throughput
- Parallel Filesystem (chkpt, mpi-io datasets)
  - Set of nodes (clients), running in parallel, read or write a a shared data set.
- To Improve Performance: Constrain the problem, study access patterns, optimize
  - TCP/IP --- MPI, Posix I/O --- MPI-IO, NFS or GFS --- PVFS
- HP: "Lustre: Could become ubiquitous storage system" (Replace NFS) ????? **AND** do Parallel I/O???

HPC Architectures Will Evolve, But The OS Will Be Called Linux

- Programmable Accelerators
  - Viz: Linux with graphic cards nearly a de-facto architecture for large scale viz
    - E.g. TeraGrid: 96 P4s with graphics cards
  - FPGA
    - "accelerated a small test kernel to over 400 times its performance on a 2 GHz Xeon processor. "
- Arrays of densely packed compute engines bolted to a Linux cluster
- PIMs inserted into a Linux cluster

Wrapping Up

- When will Linux achieve 90% of the HPC market?
- When will we be building 50,000 node Linux clusters?
- Architectures will evolve, they will run Linux
- Where is our "Stay On This Curve" roadmap?
- Barriers (mostly software)
  - Invest in software now!
  - We are not "ahead of the curve"
- Red Storm will run Linux on the compute nodes (Barney all but said it)
  - "50% of time is spent on buggy hardware", yet "custom MKL is less risky than Linux"… Huh? What about all the work AMD/SuSE/RH and their efforts

Argonne MCS Activities

- Parallel File Systems (PVFS, SDM SciDAC)
- MPI (MPICH)
- System Software (SSS SciDAC)
- Distributed Computing (TeraGrid)
- Experimental Architectures (FPGAs, Graphics)
- Scalability Test beds (Chiba City)
  - OS, Filesystem, middleware
- Numeric Libraries for accelerators (CCA SciDAC)